Time Series Prediction Based on Linear Regression and SVR*

Kunhui Lin[1,**],Qiang Lin[1], Changle Zhou[2], Junfeng Yao[1]
(1. Software School, Xiamen Univ. , Xiamen 361005, Fujian, China ;
2. Computer Science Dept., Xiamen Univ. , Xiamen 361005, Fujian, China)
khlin@xmu.edu.cn

Abstract

The application of SVR in the time series prediction is increasingly popular. Because some time series prediction based on SVR wasn't very nice in the efficiency of the forecast, this article presents a new regression based on linear regression and SVR. The new regression separates time series into linear part and nonlinear part, then predicts the two parts respectively, and finally integrates the two parts to forecast. Experiments show that the new regression advances the precision of the forecasting compared to the common SVR.

1. Introduction

As the development of economy, competition becomes more and more fierce. It is importance to forecast the social economic activities if you want to succeed in grants competition. So uncertain or unknown event in the economic activities was forecasted to get scientific forecasting, using various necessary means. That can provide scientific evidence for making economic decision to guide the economic work at present and future. Time series analysis, forecasting and digging are important manners in dynamic system modeling. These methods are widely used in meteorology, finance, medical, electricity, hydrology, industrial control etc. It deserves significant value of study [1].

There are some traditional methods in the time series prediction, such as Neural Networks [2], Statistic [3]. But they have their limitation respectively. Statistical method only fits simple time series, but this method isn't nice for complicated time series. Neural Networks has good learning capability, but it often causes underfitting and overfitting that the generalization capability becomes weaker. SVM, which is presented by Vapnik etc in 1995, is used more widely in time series at current[4-7]. Different from the traditional statistical learning theory, SVM based on the principle of structural risk minimization can solve the problem of overfitting effectively and has good generality capability and better classification accuracy. So it has a better prospect of application.

The fluctuation of economy time series is influenced by long-term tendency(T), seasonal variation(S), cycle variation(C) and irregular variation(I). Based upon additive model, time series can be represented by the sum of the above four factors.

Time series is divided to two parts in this paper. One is the stable part, which contains long-term tendency and indicates the relatively steady part of the time series; the other is the instable part, which contains seasonal variation, cycle variation and irregular variation. It indicates all kinds of factors that affect the time series to be unsteady. Common SVR neglects the inherent characteristics of time series, and can't analyze specifically of specific matters.

A new SVR is presented in this paper. It splits time series into linear part and nonlinear part, takes linear part as the common stability part, nonlinear part as the instability part, then forecast the two parts respectively, and finally integrates the two parts to get forecasting result. The regression analyzes specifically of specific matters, and enhances the efficiency of the forecasting to some extent.

2. Pre-processing of time series

When you want to forecast the linear part and nonlinear part of the time series, the pre-processing of time series must do first. Time series is split by first-order linear regression in this paper.

For a time series $\{x_t\}$, Where t=1, 2, …… , N. Do first-order linear regression, as use formula below.

$$b = \frac{\sum x_i t - n\bar{x}\bar{t}}{\sum t^2 - n\bar{t}^2} \quad (1)$$

* This work is supported by the National Hi-Tech Research and Development Program (863) of China (No. 2006AA01Z129), the National Natural Science Foundation of China (No. 60672018) and the 985 Innovation Project on Information Technique of Xiamen University (0000-X07204)
** Corresponding author: Kunhui Lin(khlin@xmu.edu.cn)

$$a = \bar{x} - b\bar{t} \tag{2}$$

Where, $\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_i$, $\bar{t} = \frac{1}{n}\sum_{t=1}^{n} t$

Form above, get the a, b Then set $x' = a + bt$. The x' is the linear part of the time series $\{x_t\}$. Set $x'' = x - x'$, the x'' is the nonlinear part of the time series $\{x_t''\}$.

3. Phase space reconstruction and SVR

3.1. Phase space reconstruction

As to forecast the linear part of time series, the dynamics movement regularity of the part must be researched. So phase space must be reconstruction.

For the series of the nonlinear part $\{x_t\}$, make $X_i = \{x_i, x_{i+1}, \cdots, x_{i+p-1}\}, Y_i = \{X_{i+p}\}$. In which, p is the embedding dimension. There exists a map, $f: R_p -> R$, let

$$Y_i = f(X_i) \tag{3}$$

So, get pairs of specimens (X_i, Y_i), and use the pairs of specimens to estimate the map f.

3.2. Support vector regression(SVR)

The principle of SVR[8]: The data Xi are mapped into a high dimensional feature space with a nonlinear mapping φ. The nonlinear problem of low dimension translates into the linear problem of high dimension. That is to ascertain the regression function below:

$$f(x) = w \cdot \varphi(x) + b \tag{4}$$

Different from the traditional statistical learning theory, SVM minimizes structural risk to find the function f. It can solve the problem of overfitting effectively and has good generality capability.

As the ε-insensitive loss function is presented by Vapnik, structural risk minimization can be figured by the formula below.

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*), \tag{5}$$

$$s.t. \quad ((w \cdot \varphi(x_i)) + b) - y_i \le \varepsilon + \xi_i, i = 1,2,\ldots,l,$$

$$y_i - ((w \cdot \varphi(x_i)) + b) \le \varepsilon + \xi_i^*, i = 1,2,\ldots,l,$$

$$\xi_i^{(*)} \ge 0, i = 1,2,\ldots,l,$$

Where , $w \in R^n, b \in R$, φ is the mapping form the input feature space to the high dimensional feature space, (*)indicates vector with mark * or not. For example $\xi_i^{(*)} \ge 0$ means $\xi_i \ge 0$ and $\xi_i^* \ge 0$.

Base on Lagrange function and dual problem, get:

$$\min_{a^{(*)} \in R^{2l}} \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)K(x_i, x_j) + \varepsilon\sum_{i=1}^{l}(\alpha_i^* + \alpha_i) - \sum_{i=1}^{l} y_i(\alpha_i^* - \alpha_i), \tag{6}$$

$$s.t. \ \sum_{i=1}^{l}(\alpha_i^* - \alpha_i) = 0,$$

$$0 \le \alpha_i^{(*)} \le C, i = 1,2,\cdots,l,$$

Where $K(x_i, x_j)$ is kernel function, the result of dual problem depends on the kernel function $K(x_i, x_j)$ $(i = 1,2,\cdots,l)$ and the constant C.

Get the result, $w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\varphi(X_i)$, b can be got by substitution of a support vector, then the function f is:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\varphi(X_i), \varphi(X)) + b \tag{7}$$

4.Linear regression and SVR mixed forecasting

4.1.Construction of the mixed forecasting model

For a given time series $\{x_t\}$,Where t=1,2,⋯⋯,N. First divide the time series into linear part $\{x_t'\}$ and nonlinear part $\{x_t''\}$. The forecasting values of linear part is $\hat{x}_t' = a + bt$.

Then forecast the nonlinear part $\{x_t''\}$ using SVR. Get first N_{train} data as training patterns of forecasting model, last N_{test} data as test patterns of forecasting model. After reconstruction of phase space, get $N_{train} - p$ pairs of specimens as training patterns, N_{test} pairs as test patterns.

Training the patterns, can get the regression, $f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\varphi(X_i), \varphi(X)) + b$. By that, we can get $\hat{x}_t''$, the forecasting value of x_t''.So overall forecasting value $\hat{x}_t = \hat{x}_t' + \hat{x}_t''$.

4.2.The choice of estimation standard of the mixed forecasting model

Statistics to the evaluate simulation effect and predictive validity of forecasting model has: Root Mean Square Error(RMSE), Mean Absolute Error(MAE) and Average Relative Error(ARE). ARE is often used for it is

directly observable. In this paper also adopt ARE, as below:

$$ARE = \frac{1}{n}\sum_{t=1}^{n}\frac{\left|y_t - \hat{y}_t\right|}{\left|y_t\right|} \quad (8)$$

Where, y_t is the actual value of time series, $\hat{y}_t$ is the forecasting value.

4.3.The choice of parameter of the mixed forecasting model

The main parameters: the embedding dimension of time series p, penalty factor C, the parameter of loss function ε, the parameter of kernel function[5].

4.3.1. The embedding dimension of time series p

The p is related to reconstruction of phase space of nonlinear system.

4.3.2. Penalty factor C

The C is the balance between model complexity and approximation error. Too high or too low, the generalization ability of model becomes weaker.

4.3.3.The parameter of loss function ε

The ε adjust the size of approximation error of the regression to control the number of the support vector and the generalization ability. The bigger of the value of ε, the lower of the precision is.

4.3.4. The parameter of kernel function

For different kernel functions, there was no significant change to the number of support vector, but the parameter of kernel function is important to the precision of forecasting. RBF kernel function is often used. This experiment also uses the RBF kernel function. As below:

$$K(x,y) = \exp\left[-\|x-y\|^2 / 2\sigma^2\right] = \exp\left[-\|x-y\|^2 / D\right] \quad (9)$$

Where, $D = 2\sigma^2$.

The choice of the four parameters above bases on minimization of ARE, the evaluation standard of the forecasting model.

5. Experiment

5.1.The environment of experiment

The software and hardware environment in this experiment is as follows:

Hardware environment: Pentium(R)4 Lenovo compatible machine, 3.00 GHz CPU, 1.00 GB memory;

OS: Microsoft Windows XP;

Programming language: Matlab 7.0;

The data of this experiment is nationwide monthly sales of automobile from Guo Yan Net and Market Research Net of China[9-10].It is important to the consumption of product oil and has predictability. Choose the data of the nationwide monthly sales of automobile from 2001.1 to 2006.6. The data of first 60 month is used to construct the model. The test data set comes from last 10 month.

5.2.The result of experiment

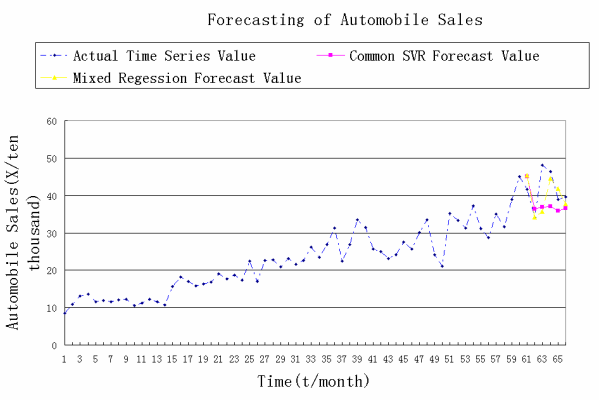

Figure 1. Comparison between two manners

Figure1 show that the mixed model is better than common SVR in result, as table1 show

Table 1 prediction data between two manners

Source data	41.5652	34.179	48.0958
SVR data	45.0366	36.2981	36.8933
New model data	45.4356	34.1775	35.7289
Source data	46.3303	38.39649	39.6433
SVR data	37.1143	35.831	36.5036
New model data	44.5543	41.7416	37.7681

Table 1 shows that the precision result of the new model is better than the SVR.

5.3.The analysis of common SVR model

5.3.1. Choice of the embedding dimension p

With ARE as evaluation standard, compare the forecasting error at different value of p, when D=0.1, ε =0.001,C=39. as table 2 below.

Table 2 Compare the forecasting at different value of P

p	4	5	6
forecasting error	0.14137	0.12283	0.1539
p	8	10	12
forecasting error	0.15807	0.20495	0.21655
p	14	16	
forecasting error	0.22845	0.22952	

Table 2 shows that the precision of forecasting is best when p is 5, so p=5.

5.3.2.Choice of the penalty factor C

Compare the forecasting error at different value of C, when D=0.1, ε =0.001, p=5. as table 3. It shows that C can be choosed from 20 to 50.

Table 3 Compare the forecasting at different value of C

C	1	10	20
Forecasting error	0.17626	0.14756	0.15105
C	30	38	39
Forecasting error	0.13263	0.12422	0.12283
C	40	41	45
Forecasting error	0.12301	0.1232	0.14273
C	50	100	1000
Forecasting error	0.14236	0.14236	0.13597

5.3.3.Choice of the parameter of loss function ε

Compare the forecasting error at different value of ε , when D=0.1, p=5,C=39. as table 4 below.

Table 4 Compare the forecasting at different value of ε

ε	0.1	0.01	0.001
Forecasting error	0.12700	0.23454	0.12283
ε	0.005	0.0001	
Forecasting error	0.14581	0.15173	

Table 4 shows that ε can choose 0.01.

5.3.4. Choice of the parameter of kernel function D(σ)

Compare the forecasting error at different value of D, when ε =0.001, p=5,C=39. as table 5 below.

Table 5 Compare the forecasting at different value of D

D	0.01	0.1	0.5
Forecasting error	0.30881	0.12283	0.24848
D	1	3	10
Forecasting error	0.16986	0.54644	0.58381

Table 5 shows that D can choose 0. 1.

When ε =0.001, p=5, C=39, D=0.1, ARE is minimal, it is 0.12283.

5.4. The analysis of mixed forecasting model

Using the ARE evaluation standard, when ε =0.01, p=12, C=20, D=1.9, ARE is minimal, it is 0.084531. Result shows that ARE of the mixed forecasting model lower than common SVR. The new regression is better than common SVR 40% in precision of forecasting.

6. Conclusions

This paper predicts the time series by a new SVR that combines linear regression. The new model separates time series into linear part and nonlinear part, and predicts the nonlinear part by SVR. Experiments show that the new regression advances the efficiency of the forecasting compared to the common SVR. However, it's just a common case to use first-order linear model as stable part. In the future work, basing on the difference characteristic, the stable part can be second-order linear or third-order linear model, so much as not to be linear, such as exponential curve model. Automatically adaptive of the parameter of the model also can be taken into consideration.

Acknowledgment

This work was supported by the National Hi-Tech Research and Development Program (863) of China (No. 2006AA01Z129), the National Natural Science Foundation of China (No. 60672018) and the 985 Innovation Project on Information Technique of Xiamen University (0000-X07204).

References

[1] Qu wenlong, Fan guanquan, Yang bingru. Complexity time series predict basing on SVR. Computer Engineering, 2005, (23).
[2] Guo zhigang. Instrument investment analysis basing on Neural Networks and genetic algorithm. South Western University of Finance and Economics, 2006.
[3] He chuan. Statistical modeling and forecasting on nation value added. Journal of Liaoning Normal University,2006.
[4] Wang peng, Zhu xiaoyan. Model choice and application of SVM basing on RBF. Computer Engineering and Applications, 2003, (24).
[5] Chen guo. SVR time series model optimization basing genetic algorithm. Chinese Journal of Scientific Instrument, 2006, (9).
[6] David Lindsay, Sian Cox. Effective Probability Forecasting for Time Series Data Using Standard Machine Learning Techniques. ICAPR 2005, LNCS 3686, pp. 35–44, 2005.
[7] Xiaodong Wang, Haoran Zhang. Time Series Prediction Using LS-SVM with Particle Swarm Optimization. ISNN 2006, LNCS 3972, pp.747 – 752, 2006.
[8] Nello Cristianini,John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Publishing House of Electronics Industry.2000.
[9] Guo Yan Net.,2006, http://www.drcnet.com.cn,
[10] Market Research Net of China, 2006.